International Journal of Languages' Education and Teaching

E-ISSN: 2148-2705

Volume 11, Issue 3, September 2023

Original Research Article

Statistical Analysis of the Effectiveness of An Augmented Reality Language Learning Tool for Pre-elementary School Children

Elif TOPSAKAL [1] *& Oğuzhan TOPSAKAL* [2]

Abstract

Augmented Reality (AR) technology has been utilized in many fields, including language education. However, there is limited research on its application in early childhood education. This paper focuses on the statistical analysis of an AR application's efficacy for learning new vocabulary at the pre-elementary school level. We used a pre-test/post-test experimental design study to compare the effectiveness of learning foreign language vocabulary using traditional approaches (flashcards) and using an AR app. The data were collected from three daycares on students at three age groups, 4, 5, and 6 years old. The Mixed ANOVA Model is used to compare the control and treatment groups' means through pre-test and post-test scores. We also conducted a reliability analysis to see groups' pre-test and post-test scores are internally consistent. Cronbach's alpha scores were used as the reliability estimate. Moreover, the item-total correlation and item difficulty for each age group were calculated. The results show that Cronbach's alpha values for all age groups are between .80 and .95. (>.70). The item analysis results identified the easiest and the most difficult questions on the vocabulary test. The results of the Mixed ANOVA Model show the statistically significant difference between pre-tests and post-test results (time factor) for each age group (p=0.00, p<0.05 for each age group). We also show there is an interaction effect between pre/post-test scores and control/treatment groups for four-year-olds (p=0.00, p<0.005) and five-year-olds (p=0.045, p<0.05) in learning a foreign language when AR apps are utilized in the classroom.

Keywords

Augmented Reality

Foreign Language Learning

Early Education

Mixed ANOVA

Cronbach's alpha

Article Info

Received	19.07.2023
Reviewed	30.08.2023
Published	30.09.2023
Doi Number	10.29228/ijlet.71198

Reference

Topsakal, E., & Tospakal, O. (2023). Statistical analysis of the effectiveness of an augmented reality language-learning tool for pre-elementary schoolchildren. *International Journal of Languages' Education and Teaching*, 11(3), 111-127.

[1] Ph.D. Candidate, University of South Florida, elift@usf.edu, ORCID: 0000-0001-6651-669X

[2] Assistant Professor, Florida Polytechnic University, otopsakal@floridapoly.edu, ORCID: 0000-0002-9731-6946

International Journal of Languages' Education and Teaching

E-ISSN: 2148-2705

Cilt 11, Sayı 3, Eylül 2023

Orijinal Araştırma Makalesi

İlkokul Öncesi Çocuklar İçin Artırılmış Gerçeklikli Dil Öğrenme Aracının Etkinliğinin İstatistiksel Analizi

Özet

Artırılmış Gerçeklik (AR) teknolojisi, dil eğitimi de dâhil olmak üzere pek çok alanda yaygın bir şekilde kullanılmaktadır. Ancak, AR teknolojisinin erken çocukluk eğitimindeki uygulamaları üzerine sınırlı araştırma bulunmaktadır. Bu makale, anaokulu öncesi dönemi çocuklarında İngilizce yeni kelime öğrenme süreçlerinde artırılmış gerçeklik (AR) uygulamasının etkisini incelemeyi amaçlamaktadır. Geleneksel yöntemlerle (örneğin, resimli kelime kartları) kelime öğrenme ile AR uygulaması kullanarak öğrenmenin etkinliğini karşılaştırmak için ön-test/son-test deneysel tasarımına dayalı bir çalışma yürütülmüştür. Veriler, üç farklı anaokulundan yaşları 4, 5 ve 6 olan öğrencilerden toplanmıştır. Deney gruplarında öğretim materyali olarak AR uygulaması kullanılırken, kontrol gruplarında resimli flascardlar kullanılmıştır. Kontrol ve deney gruplarının ön-test ve son-test puanlarını karşılaştırmak amacıyla Mixed ANOVA modeli kullanılmıştır. Aynı zamanda grupların ön-test ve son-test puanlarının iç tutarlılığını değerlendirmek için güvenilirlik analizi gerçekleştirilmiştir. Bu analizde Cronbach's alfa katsayısı kullanılarak güvenilirlik tahminleri elde edilmiştir Ayrıca her yaş grubu için hangi kelimelerin daha zor hangi kelimelerin daha kolay öğrenildiğinin analizi yapılmıştır. Sonuçlara göre, tüm yaş grupları için Cronbach alfa değerleri .80 ile .95 arasında çıkmıştır (> .70). Ayrıca, her yaş grubu için ön-test ve son-test sonuçları arasında istatistiksel olarak anlamlı bir fark olduğu bulunmuştur (her yaş grubu için p<0.05). Ek olarak, dört yaş (p=0.00, p<0.005) ve beş yaş grupları (p=0.045, p<0.05) için AR uygulaması öğretim materyali olarak kullanıldığında, deney gruplarında öğrenilen kelime sayısının kontrol gruplarında öğrenilen kelime sayısına oranla daha fazla olduğu tespit edilmiştir.

Anahtar Kelimeler

Arttırılmış Gerçeklik

Yabanci Dil Öğrenimi

Erken Eğitim

Karışık ANOVA

Cronbach Alfa

Article Info

Gönderim Tarihi	19.07.2023
Kabul Tarihi	30.08.2023
Yayın Tarihi	30.09.2023
Doi Numarası	19.07.2023

Kaynakça

Topsakal, E., & Topsakal, O. (2023). İlkokul öncesi çocuklar için artırılmış gerçeklikli dil öğrenme aracının etkinliğinin istatistiksel analizi. *International Journal of Languages' Education and Teaching*, 11(3), 111-127.

Introduction

It is vital to motivate the learner, increase their interest and encourage them through engaging activities to teach the foreign language effectively, especially if the learner is a child. (Gundogmus, & Orhan, 2016; Chang et al., 2011; Scrivner et al., 2017; Vate & Lan, 2012). Augmented reality (AR) technology can be utilized to grab children's attention and make them engaged in foreign language learning activities.

The studies about teaching a foreign language were mostly performed on students who are in elementary (Barreira et al., 2012; Chang et al., 2011; Solak & Cakir, 2017), secondary (Gundogmus & Orhan, 2016; Kucuk et al., 2014), and college levels (Ibrahim et al., 2018; Kayaoğlu et al., 2011). Only a few studies test AR's effectiveness on pre-elementary school children (Dalim et al., 2017; Chen & Chan, 2019). The reason might be that data collection from young children is challenging (James & Christensen, 2008; Malet et al., 2010). This study endeavors to bridge the existing gap in literature by offering insights into the efficacy of augmented reality (AR) as an instructional tool for teaching foreign languages to young children.

In this research, an experimental study has been conducted to see the effects of AR technology on teaching English to young children and the treatment groups were compared with control groups regarding the number of words they learned. A pre-test was administered to assess the English level of groups at the beginning of the study. At the end of the teaching sessions, we applied the same test used in the pre-test as the post-test to both experimental and control groups to assess how much each group improved their English vocabulary. It is seen that children using the AR app learned English more effectively and achieved better scores in the post-test than the post-test scores of children who studied English using the conventional methods (flashcards). While the children learning through the traditional techniques increased their vocabulary 25%, 17%, and 24% for 4, 5, and 6 age levels, children utilizing the AR app increased their vocabulary by 78%, 38%, and 73%.

This study performs statistical analysis and provides the reliability and item analysis results for the pre-test and post-test results of the experimental research. A statistically significant mean difference was shown between the group results using a mixed ANOVA model. We believe this study can provide guidance and examples for researchers working on foreign language education and willing to apply statistical methods to analyze their results.

Literature Review

AR technology has been utilized in many fields, including tourism, advertisement, training, military, medicine, and education (Carmigniani & Furht, 2011). AR can help to produce positive outcomes in education (Pellas et al., 2018; Saidin et al., 2015). There have been studies to see the effect of AR on teaching a foreign language to students (Salmon & Nyhan, 2013; Scrivner et al., 2017; Vate & Lan, 2016; Yilmaz, 2016).

In a similar study, Chen and Chan explored AR's potential for language learning for children by comparing conventional approaches and could not find a significant difference. The insignificant difference might be due to the similarity between the learning activities (Chen & Chan, 2019). In the recent study, despite employing the same instructional plan, a noteworthy distinction emerged between the experimental and control groups. The use of an AR app as the learning tool led to a significant variance in vocabulary acquisition.

Motivation is one of the most critical factors facilitating learning, especially in foreign language education (Gundogmus et al, 2016; Kucuk et al., 2014; Salmon & Nyhan, 2013; Solak & Cakir, 2017). According to the results of Solak and Cakir study (2017), the use of AR technology provided a more effective environment for vocabulary learning than traditional methods by increasing learners' performance. Gundogmus et al. (2016) collected data from 60 students in a secondary school by using 15 items of the "Augmented Reality Applications Attitude Scale in Secondary Schools" scale. According to the study results, the students who use AR applications in language learning had a positive attitude towards the mobile AR application and AR attracted their attention and increase their motivation for learning by providing more enjoyable learning sessions. Kucuk et al. (2014), examined the achievement, attitude, and cognitive load levels of students in learning English by Augmented Reality (AR). They found that secondary school students had a low anxiety level while learning English with the aid of AR and were willing to use such applications in their feature courses. Also, the study indicated that students who have positive attitudes towards AR applications were more successful compared to the other students.

Games and surprise factors are utilized in early childhood education at pre-elementary school ages to motivate. AR technology is astonishing to young children, and it is like magic as it brings virtual objects to existence (Barreira et al., 2012; Dalim et al., 2017; Yilmaz, 2016). According to the case study of Barreira et al. (2012), children who used Augmented Reality games (MOW-Matching Objects and Words) showed better progress than those who used only traditional methods while learning new language, and "the use of AR games has a positive pedagogical impact in the learning process concerning young children, more exactly in the progressive domain of oral recognition of words and concepts and their corresponding written form" (p.6). Dalim et al. (2017) used an Augmented reality (AR) tool (TeachAR), to teach colors, shapes, and prepositions in English to children who are not native English speakers. Their comparison study (comparison learning with AR with learning with traditional methods) indicated that children who learned with TeachAR system had a better learning outcome than the traditional system. Additionally, children had a more enjoyable time while using AR-based methods.

Research Methods

Method

In the recent study, pre-post test experimental study design was used. The control and experiment groups were created according to the pre-test results of children to increase the heterogeneity within the groups and homogeneity between the groups.

Participants, the AR App, and the Teaching Method

The study was conducted on pre-elementary school children at ages 4, 5, and 6 in three daycares in Bursa, Turkey. There was a total of 85 participants. Twenty-one of these children were in the four-year-old group, twenty-four were in the five-year-old group, and forty were in the six-year-old group. We put the students into experimental and control groups randomly. While the control group learned English with conventional flashcards, the experimental group learned English vocabulary utilizing a language learning tool. The tool includes 40 image cards (images of animals), 60 word cards (written text of animal names and action words such as walk, run, jump, etc.), and an AR app that works with the cards. The AR app provided an interactive and engaging environment through games and fun interactions to teach vocabulary about animal names and the actions (walk, run, fly, etc.) that an animal can perform.

During the teaching sessions, both control and experimental groups spent three class periods learning English animal names and action words. For 4-, 5-, and 6-years old experiment and control groups, 15, 20, and 25 animal names were taught, respectively. With the experimental group, the teacher first showed how the AR app works. Then, in small groups, children started to play with the AR app. When an animal picture is shown to the Android tablet camera, the App pops up the 3D model of the animal, and then the audio is played pronouncing the name of the animal in English. With the control group that utilizes the flashcards, the teacher first showed pictures of each animal and repeated its name three times. Then, the teacher let each student say the animal names three times.

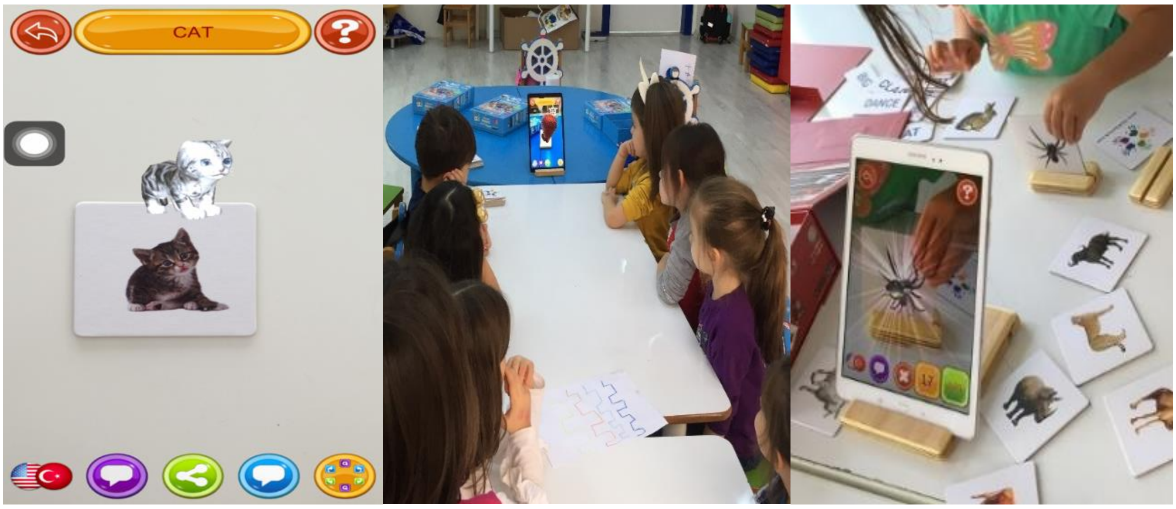

Figure 1. *On the left a screenshot of the AR app; in the middle, children are listening to the instructions about the AR app, and on the right, a child is using the AR app.*

Measurement Instrument

The measurement instrument is designed as a pictured English vocabulary test and includes questions about the animal names and action words that an animal can perform. The action-words and animals were selected by reviewing several pre-elementary school English learning applications and books. The items were asked to students by showing the animals' pictures and asking the English correspondents. We used the same picture of animals used on the flashcards and in the AR app. We asked the questions in Turkish (i.e., "How do you say 'kedi (cat)' in English?", "How do you say 'zıpla (jump)' in English?"). Each age group had a different number of items; 15 animal names and three action words for four-year-olds, 20 animal names, and six action words for five-year-olds, and 25 animal names and ten action words for six-year-olds. The items used for age groups 4, 5, and 6 are listed in Tables 2, 3, and 4, respectively.

Methodology for the Reliability Analysis of Test Scores

Classical Test Theory (CTT) is one of the most common approaches used to measure test scores changes (Frey, 2017). There are four major statistics reported in the framework of the CTT; "1) Item Difficulty, 2) Item-Test Correlation, 3) Reliability Coefficient, 4) Standard Error of Measurement (SEM)" (Zeng & Wyse, 2009). Our study used the "internal structure analysis" method, which is one of the primary methods for estimating reliability coefficients.

We utilized Cronbach's alpha as a reliability estimate for pre-elementary school students (4, 5, and 6-year-olds). We used the same test as pre-test and post-test and reached four different reliability estimates for each age group (pre-test reliability for control and treatment groups and post-test reliability for control and treatment groups). According to George & Mallery (2003), reliability estimate above 0.9 is considered as excellent, and above 0.8 as good reliability.

Methodology for Comparing Group Means

We used Mixed ANOVA Model (Repeated Measures ANOVA) as the statistical method to analyze the comparison of control and treatment groups' means through pre-test and post-test scores. We conducted three different Mixed ANOVA models for each age group and tested three different null hypotheses in this study. These hypotheses are as follows.

- There is no significant difference between pre-test and post-test results.
- There is no significant difference between the control and treatment groups of test scores' means.
- There is no interaction effect between the times (pre-test/post-test) and groups (control/treatment)

We used Mixed ANOVA Design because we have repeated measures (nested data). We used the Huyn-Feldt (HF) adjusted p-values to report inferential statistics. We also decided to set our alpha value at .05 even though we tested three hypotheses for each age group because our sample size is considerably small.

Methodology for the Item Difficulty/Variance Analysis of Pre-test and Post-test Results

"The mean of a dichotomous item is equal to the proportion of individuals who endorsed/passed the item" (Kline, 2005, p. 96). Item difficulty is represented by "p" and ranges between 0 and 1. If the value of an item is close to zero, it means the item is difficult. If it is close to 1, it means that the item is easy to answer for the respondents. This analysis gives us very useful information for designing tests of ability or achievement. The items with a p-value of 1.00 or 0.00 are useless because they do not differentiate between individuals. Besides, the p-value of 0.50, which means 50% of the group correctly answered the item, provides the highest differentiation levels between individuals in a group (Kline, 2005; Sonepad, 2014).

The variance of a dichotomously scored item is the product of p, the proportion of individuals who answer the item correctly, and q, which is the proportion of individuals who answered the item incorrectly. The item variance equals p x q and gives us the differentiation made by that item among the respondents. That means each person who answered the item correctly is differentiated from the one who answered incorrectly. We calculated each age group's item difficulty and item variance for both pre-test and post-test scores for all control and treatment groups.

Methodology for Finding the Item-Test Correlations

Item-test correlation is calculated using the Pearson Correlation Coefficient, which shows the correlation between scores where one item of each pair is an item score, and the other is the total test score. The greater value of the coefficient indicates a stronger correlation between the test items and the total test and increases the test's internal consistency (Salkind, 2010). Higher item-test correlation also indicates that high ability examinees tend to get the item correct, and low ability examinees tend to get the item incorrect (Zeng & Wyse, 2009).

Higher positive values for the item-total correlation shows that the item is a strong item for discriminating the high and low performing participants. Negative values mean the opposite; low performing participants are more likely to get the item correct.

Results

We used SPSS Data Analysis program version 25 for statistical analysis. In the following subsections, we present the results for the reliability analysis, item difficulty/variance analysis, item-test correlations, and the comparison of the control and treatment group's means.

Reliability Analysis

As shown in Table 1, the reliability estimates range from .92 to .97 for four-years-old, from .88 to .94 for five-years-olds, and from .80 to .94 for six-years-olds. While the reliability of four-years-olds' pre-test-treatment group scores has the highest value, the reliability of six-years-olds' post-test-control group scores has the lowest value. According to our result, the Cronbach's alpha values for all age groups are between .80 and .95. (>.70) which shows good reliability (Tavakol & Dennick, 2011).

Table 1. *The results of the reliability analysis for 4-years-olds.*

Age	Test	Group	Removed Items (Zero Variance)	Number of items	Cronbach's Alpha
4 Years (n=21)	Pre-test	Control	4,6,7,8,9,14	12	.96
		Treatment	6,8,9	15	.97
	Post-test	Control	4.6.7.8	14	.95
		Treatment	-------------------	18	.92
5 Years (n=24)	Pre-test	Control	6,8,10,19,20,21,26	19	.93
		Treatment	8,11	24	.88
	Post-test	Control	6,8,10,19,20,26	20	.93
		Treatment	8	25	.94
6 Years (n=40)	Pre-test	Control	22,23,24,33	31	.94
		Treatment	6,23,24,25	30	.95
	Post-test	Control	3,21,24,33	31	.80
		Treatment	3,21	33	.92

Comparing Group Means (Mixed ANOVA Model's Results)

Table 2 shows that the mean differences of the pre-test and post-test results for the treatment group are greater than the control group for four-year-olds (1.72> .20), five-year-olds (2.69>1.36), and six-year-olds (6.78> 4.11).

Table 2. *Descriptive statistical information for 4-, 5-, and 6-year-olds.*

4 years old (n=21)	Pre-test scores		Post-test scores	
Group	Control	Treatment	Control	Treatment
Mean(M)	1.50	2.64	1.70	4.36
SD	3.47	4.80	3.87	4.98
Min	0.00	0.00	0.00	1.00
Max	11.00	15.00	12.00	17.00
Skewness	2.77	2.13	2.64	1.99
Kurtosis	7.95	4.29	7.15	3.87

5 years old (n=24)	Pre-test scores		Post-test scores	
Group	Control	Treatment	Control	Treatment
Mean(M)	5.00	6.54	6.36	9.23
SD	5.21	6.56	5.53	6.72
Min	0.00	0.00	1.00	2.00
Max	16.00	20.00	18.00	23.00
Skewness	0.93	0.74	0.96	0.73
Kurtosis	0.25	- 0.45	0.20	- 0.37

6 years old (n=40)	Pre-test scores		Post-test scores	
Group	Control	Treatment	Control	Treatment
Mean(M)	5.28	5.95	9.39	12.73
SD	5.06	6.94	4.34	6.91
Min	0.00	0.00	1.00	4.00
Max	18.00	24.00	18.00	29.00
Skewness	1.04	1.52	0.15	0.93
Kurtosis	0.70	1.67	-0.20	0.73

Figure 2 presents the mean changes of pre/post-test score for the treatment and control groups. The figure shows that mean changes for the treatment groups are more significant than the mean change for the control groups for four, five, and six-year-olds.

Since we conducted three analyses for each age group, we present the results of the repeated ANOVA statistics separately in Tables 3, 4, and 5. The 'Time' row represents the pre-test and post-test, the 'Group' row represents control and treatment groups, and the 'Time*Group' row represents the interaction effect between the time and group in Tables 3, 4, and 5. We used the alpha level p=0.05 for significance.

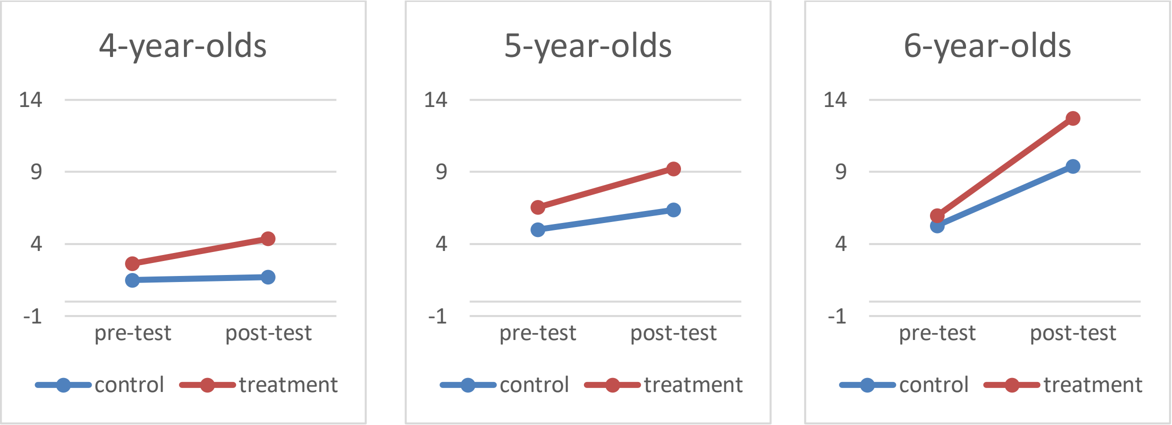

Figure 2. *The changing of means for 4-, 5-, and 6-year-olds control-treatment groups.*

The results for 4-year-olds in Table 3 shows that there is an interaction effect between within-subject (time) factors and between-subject (group) factors (F (1,19) = 19.70, pHF=.00, p<.05). That means that the test results depend on the group factor. Besides, there is a statistically significant difference between pre-test and post-test results for both groups (F (1,19) =31.38, pHF=.00, p<.05), but there is no significant difference between control and treatment groups (F (1,19) = 1.01, p=.33, p>.05). It means that the groups' results changed the same amount over time. The correlation between time 1 (pre-test) and time 2 (post-test) is very high according to paired sample test results (n=21, r=.973). Our data is also consistent with the sphericity assumption (epsilon HF= 1.00).

Table 3. *Results of the repeated ANOVA statistics for 4-year-olds.*

Effect	SS	df	MS	F	p-value	pHF	Epsilon HF
Group	18.910	1	18.910	1.007	.328		
Within	1612.427	22	73.292				
Time	9.728	1	9.728	31.376	.000	.000	1.000
Time*Group	6.109	1	6.109	19.704	.000	.000	
Residual	5.891	19	.310				

Notes: *SS= Sum of Squares, *df=Degrees of Freedom, *MS=Mean Squares. * pHF= Huyn-Feldt(HF) p-value adjustment, pFH=1.0, * Epsilon HF=Huyn-Feldt sphericity parameter

Table 4 shows that there is an interaction effect between within-subject (time) factors and between-subject (group) factors (F (1,22) = 4.51, pHF=.00, p<.05) for five-year-olds. We can conclude that the test results depend on the group factor. Besides, there is a statistically significant difference between pre-test and post-test results for both groups (F (1,22) =42.02, pHF=.00, p<.05) but again, there is not any significant difference between control and treatment groups (F (1,22) = .789, p=.38, p>.05). The correlation between time 1(pre-test) and time 2(post-test) is very high according to paired sample test results (n=24, r=.965). Five years-old data is also consistent with the sphericity assumption (epsilon HF= 1.00).

Table 4. *Results of the repeated ANOVA statistics for 5-year-Olds.*

Effect	SS	df	MS	F	p-value	pHF	Epsilon HF
Group	57.823	1	57.823	.789	.384		
Within	1612.427	22	73.292				
Time	49.009	1	49.009	42.023	.000	.000	1.000
Time*Group	5.259	1	5.259	4.510	.045	.045	
Residual	25.657	22	1.166				

Table 5 shows the results of the Repeated ANOVA Statistics for 6-Year-Olds. Different from other age groups, there is no interaction effect between within-subject (time) factors and between-subject (group) factors (F (1,38) = 3.09, pHF=.87). That means that the test results do not depend on the group factor. Also, there is no significant difference between control and treatment groups (F (1,38) = 1.30, p=.26). It means that the control and treatment groups' results changed the same amount over time. However, there is a statistically significant difference between pre-test and post-test results for both groups (F (1,38) =51.72, pHF=.00, p<.05). Furthermore, there is a correlation between time 1 (pre-test) and time 2 (post-test) according to paired sample test results (n=40, r=.677), and our six-year-old data is also consistent with the sphericity assumption (epsilon HF= 1.00).

Table 5. *Results of the repeated ANOVA statistics for 6-year-olds.*

Effect	SS	df	MS	F	p-value	pHF	Epsilon HF
Group	79.801	1	79.801	1.296	.262		
Within	2340.386	38	61.589				
Time	586.367	1	586.367	51.720	.000	.000	1.000
Time*Group	35.067	1	35.067	3.093	.087	.087	
Residual	430.801	38	11.337				

Notes: **SS= Sum of Squares, *df=Degrees of Freedom, *MS=Mean Squares. * pHF= Huyn-Feldt(HF) p-value adjustment, pFH=1.0, * Epsilon HF=Huyn-Feldt sphericity parameter*

Item Difficulty and Item Variance Analysis

We present the results for item difficulty and item variance analysis for four-year-olds in Table 6. Generally, all items are difficult items for the four-year-olds' control group. However, the analysis shows that "Lion", "Wolf", "Rabbit", "Seagull", "Fox", and "Cow" are the most difficult items for the pre-test and "Lion", "Wolf", "Rabbit", "Seagull", "Fox", are the most difficult items for the post-test for the children in the control group.

For the treatment group, "Cat" has medium level difficulty as an item for pre-test, but it is an easy item for the post-test. "Wolf", "Seagull", and "Fox" are the most difficult items for the pre-test. "Wolf", "Seagull", "Fox" and, "Cow" are the most difficult items for the post-test for the treatment group.

Table 6. *Item difficulty and item variance analysis for 4-year-olds.*

		Pre-test				Post-test			
No	Item	Treatment Group (n=11)		Control Group (n=9)		Treatment Group (n=11)		Control Group (n=9)	
No	Item	(p)	(p*q)	(p)	(p*q)	(p)	(p*q)	(p)	(p*q)
1	Bear	0.18	0.15	0.11	0.10	0.36	0.23	0.22	0.17
2	Elephant	0.27	0.20	0.11	0.10	0.36	0.23	0.11	0.10
3	Cat	0.45	0.25	0.22	0.17	0.72	0.20	0.22	0.17
4	Lion	0.18	0.15	0.00	0.00	0.18	0.15	0.00	0.00
5	Duck	0.09	0.08	0.22	0.17	0.36	0.23	0.22	0.17
6	Wolf	0.00	0.00	0.00	0.00	0.09	0.08	0.00	0.00
7	Rabbit	0.09	0.08	0.00	0.00	0.18	0.15	0.00	0.00
8	Seagull	0.00	0.00	0.00	0.00	0.09	0.08	0.00	0.00
9	Fox	0.00	0.00	0.00	0.00	0.09	0.08	0.11	0.17
10	Zebra	0.09	0.08	0.11	0.10	0.18	0.15	0.11	0.17
11	Giraffe	0.18	0.15	0.11	0.10	0.18	0.15	0.11	0.17
12	Dog	0.18	0.15	0.22	0.17	0.36	0.23	0.22	0.17
13	Roster	0.18	0.15	0.11	0.10	0.27	0.20	0.11	0.17
14	Cow	0.09	0.08	0.00	0.00	0.09	0.08	0.11	0.17
15	Tiger	0.09	0.08	0.11	0.10	0.27	0.20	0.11	0.17
16	Walk	0.18	0.15	0.11	0.10	0.18	0.15	0.11	0.17
17	Run	0.18	0.15	0.11	0.10	0.18	0.15	0.11	0.17
18	Jump	0.09	0.08	0.11	0.10	0.18	0.15	0.11	0.17

*Note: p=item difficulty, p*q=item variance*

We present the results for five-year-olds in Table 7. According to the item difficulty and item variance analysis of the control group, "Spider" is the easiest item for pre-test and post-test. "Wolf", "Seagull", "Zebra", "Eagle", "Chicken" and "Wave" are the most difficult items for both pre-test and post-test.

Table 7. *Item difficulty and item variance analysis for 5-year-olds.*

		Pre-test				Post-test			
		Treatment Group (n=13)		Control Group (n=11)		Treatment Group (n=13)		Control Group (n=11)	
No	Item	(p)	(p*q)	(p)	(p*q)	(p)	(p*q)	(p)	(p*q)
1	Bear	0.15	0.13	0.09	0.08	0.23	0.18	0.18	0.15
2	Elephant	0.38	0.24	0.36	0.19	0.46	0.25	0.45	0.25
3	Cat	0.61	0.24	0.64	0.23	0.69	0.21	0.72	0.20
4	Lion	0.38	0.24	0.36	0.19	0.61	0.24	0.36	0.19
5	Duck	0.38	0.24	0.36	0.19	0.61	0.24	0.45	0.25
6	Wolf	0.08	0.07	0.00	0.00	0.15	0.13	0.00	0.00
7	Rabbit	0.38	0.24	0.36	0.19	0.46	0.25	0.36	0.19
8	Seagull	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	Fox	0.31	0.21	0.27	0.20	0.46	0.25	0.45	0.25
10	Zebra	0.38	0.24	0.00	0.00	0.38	0.24	0.00	0.00
11	Giraffe	0.00	0.00	0.09	0.08	0.08	0.07	0.09	0.08
12	Dog	0.61	0.24	0.45	0.25	0.69	0.21	0.72	0.20
13	Roster	0.15	0.13	0.18	0.15	0.23	0.18	0.18	0.15

14	Cow	0.38	0.24	0.09	0.08	0.46	0.25	0.18	0.15
15	Tiger	0.38	0.24	0.36	0.19	0.38	0.24	0.36	0.19
16	Butterfly	0.15	0.13	0.09	0.08	0.15	0.13	0.09	0.08
17	Spider	0.61	0.24	0.72	0.20	0.69	0.21	0.91	0.08
18	Dinosaur	0.38	0.24	0.09	0.08	0.15	0.13	0.09	0.08
19	Eagle	0.38	0.24	0.00	0.00	0.08	0.07	0.00	0.00
20	Chicken	0.15	0.13	0.00	0.00	0.23	0.18	0.00	0.00
21	Walk	0.15	0.13	0.00	0.00	0.23	0.18	0.09	0.08
22	Run	0.38	0.24	0.09	0.08	0.54	0.25	0.09	0.08
23	Jump	10	0.23	0.18	0.15	0.31	0.21	0.45	0.25
24	Big	0.15	0.13	0.09	0.08	0.31	0.21	0.09	0.08
25	Turn	0.15	0.13	0.09	0.08	0.23	0.18	0.09	0.08
26	Wave	0.15	0.13	0.00	0.00	0.15	0.13	0.00	0.00

*Note: p=item difficulty, p*q=item variance,*

For the treatment group, "Spider" and "Cat" are the easiest items for the pre-test, while "Cat", "Dog", "Spider" are the easiest items for the post-test. "Seagull" is the most difficult item for both pre-test and post-test. "Giraffe" is one of the most difficult items for the pre-test, but not for the post-test for the treatment group.

Table 8 shows the results for six-years-olds. For the control group, "Cat" is the easiest item for both pre-test and post-test. "Spider" is the second easiest item for the post-test. "Sparrow", "Camel", "Dragon", "Buffalo" and, "Bark" are the most difficult items for the pre-test. "Sparrow" and "Bark" are still the most difficult items for the post-test.

For the treatment group, "Cat" is the easiest item for both pre-test and post-test. "Dog" is the second easiest items for the post-test. "Wolf", "Sparrow", "Camel", "Dragon" and, "Buffalo" are the most difficult items for the pre-test. "Sparrow" is still the most difficult items for the post-test.

Table 8. *Item difficulty and item variance analysis for 6-year-olds.*

		Pre-test				Post-test			
		Treatment Group (n=21)		Control Group (n=19)		Treatment Group (n=21)		Control Group (n=19)	
No	Item	(p)	(p*q)	(p)	(p*q)	(p)	(p*q)	(p)	(p*q)
1	Bear	0.19	0.15	0.29	0.21	0.43	0.24	0.47	0.25
2	Elephant	0.52	0.25	0.53	0.25	0.71	0.20	0.65	0.23
3	Cat	0.81	0.19	0.94	0.05	1.00	0.00	1.00	0.00
4	Lion	0.43	0.24	0.59	0.24	0.76	0.18	0.70	0.21
5	Duck	0.48	0.25	0.53	0.25	0.57	0.24	0.53	0.25
6	Wolf	0.00	0.00	0.13	0.10	0.09	0.09	0.13	0.10
7	Rabbit	0.43	0.24	0.41	0.24	0.57	0.24	0.53	0.25
8	Seagull	0.05	0.04	0.05	0.05	0.05	0.04	0.05	0.05
9	Fox	0.19	0.15	0.35	0.23	0.38	0.24	0.41	0.24
10	Zebra	0.43	0.24	0.41	0.24	0.71	0.20	0.59	0.24
11	Giraffe	0.14	0.12	0.05	0.05	0.33	0.22	0.13	0.10
12	Dog	0.66	0.22	0.65	0.23	0.90	0.09	0.82	0.19
13	Roster	0.19	0.15	0.29	0.21	0.33	0.22	0.53	0.25
14	Cow	0.19	0.15	0.18	0.14	0.43	0.24	0.41	0.24

15	Tiger	0.48	0.25	0.41	0.24	0.48	0.25	0.53	0.25
16	Butterfly	0.19	0.15	0.29	0.21	0.33	0.22	0.41	0.24
17	Spider	0.62	0.24	0.65	0.23	0.81	0.19	0.94	0.05
18	Dinosaur	0.24	0.18	0.23	0.18	0.52	0.25	0.41	0.24
19	Eagle	0.05	0.04	0.05	0.05	0.05	0.04	0.05	0.05
20	Chicken	0.14	0.12	0.18	0.14	0.24	0.18	0.29	0.21
21	Sparrow	0.00	0.00	0.05	0.05	0.00	0.00	0.00	0.00
22	Camel	0.00	0.00	0.00	0.00	0.38	0.24	0.18	0.14
23	Dragon	0.00	0.00	0.00	0.00	0.33	0.22	0.00	0.00
24	Buffalo	0.00	0.00	0.00	0.00	0.05	0.04	0.31	0.21
25	Crocodile	0.05	0.04	0.18	0.14	0.14	0.12	0.23	0.18
26	Walk	0.24	0.18	0.35	0.23	0.48	0.25	0.41	0.14
27	Run	0.29	0.20	0.29	0.21	0.38	0.24	0.29	0.21
28	Jump	0.62	0.24	0.70	0.21	0.81	0.19	0.76	0.14
29	Big	0.29	0.20	0.35	0.23	0.38	0.24	0.41	0.24
30	Turn	0.33	0.22	0.35	0.23	0.43	0.24	0.47	0.25
31	Wave	0.14	0.12	0.18	0.14	0.14	0.12	0.23	0.18
32	Push	0.05	0.04	0.05	0.05	0.24	0.18	0.13	0.10
33	Bark	0.05	0.04	0.00	0.00	0.05	0.04	0.00	0.00
34	Small	0.24	0.18	0.35	0.23	0.38	0.24	0.41	0.14
35	Dance	0.29	0.20	0.35	0.23	0.48	0.25	0.35	0.23

*Note: p=item difficulty, p*q=item variance*

Item-Test Correlations

We present item-test correlation results based on the Pearson Correlation Coefficient in Table 9. The item name that corresponds to the item numbers in Table 9 can be found in Tables 2, 3, and 4. If the correlation score is higher than 0.7, that means the correlation is very strong; if the correlation score is higher than 0.4, that means the correlation is strong. Most items` scores in the tests we used for our study strongly correlate to total test scores.

When the correlation score is near zero, it means the item is very weak for discriminating the high and low-performing participants. In other words, it means that no matter their total score, all participants have similar probabilities of answering the item correctly (Fossey, 2013). According to the item-test correlation results, some test items are negatively correlated for the particular age and test groups as being the weakest items. The weak items are listed in the 'Weakest Items' column in Table 9.

Methodology for the Item Difficulty/Variance Analysis of Pre-test and Post-test Results

"The mean of a dichotomous item is equal to the proportion of individuals who endorsed/passed the item" (Kline, 2005, p. 96). Item difficulty is represented by "p" and ranges between 0 and 1. If the value of an item is close to zero, it means the item is difficult. If it is close to 1, it means that the item is easy to answer for the respondents. This analysis gives us very useful information for designing tests of ability or achievement. The items with a p-value of 1.00 or 0.00 are useless because they do not differentiate between individuals. Besides, the p-value of 0.50, which means 50% of the group correctly answered the item, provides the highest differentiation levels between individuals in a group (Kline, 2005; Sonepad, 2014).

The variance of a dichotomously scored item is the product of p, the proportion of individuals who answer the item correctly, and q, which is the proportion of individuals who answered the item incorrectly. The item variance equals p x q and gives us the differentiation made by that item among the respondents. That means each person who answered the item correctly is differentiated from the one who answered incorrectly. We calculated each age group's item difficulty and item variance for both pre-test and post-test scores for all control and treatment groups.

Table 9. *Distribution of the test items according to Item-Test (item-total) correlations.*

Age	Test	Group	Very Strong Items	Strong Items	Moderate Items	Weak Items	Weakest Items
4	pre	*Control*	1,2,3,4,5,10,12, 13,15,16,17,18				11(neg)
		Treatment	1,2,5,7,10,11 13,15,16,17,18	3, 4, 12			
	post	*Control*	2,3,5,10,12 13, 14,15,16,17,18	1			9, 11(neg)
		Treatment	2,6,9,11,14 15,16,17	1, 4, 7, 10, 13, 18	3, 12	8	5
5	pre	*Control*	4,5	1, 2, 3, 7, 11, 12, 14, 15, 16, 17, 18, 23, 24, 25	9	13, 22	
		Treatment	1,6,7,10,14 16,18,20,24 25,26	2, 3, 5, 12, 15, 17, 19, 21, 22, 23			4, 9, 13
	post	*Control*	2,4,5,9 14,15,23	3, 7, 11, 12, 16, 18, 21, 24, 25		1, 13, 17, 22	
		Treatment	2,7,9,14,16 24,25,26	5, 6, 11, 12, 13, 15, 18, 20, 21, 22, 23	1, 3, 4, 17, 19	10	
6	pre	*Control*	1,2,4,18,25 26,27,28,29 30,31,34,35	5, 7, 10, 12, 16, 17, 19, 20	6, 13, 32	3, 14, 21	8, 9, 11, 15
		Treatment	1,2,5,7,10,18 26,27,28,29 34,35	4, 11, 12, 14, 15, 16, 17, 19, 20, 30, 31	25	9,13,32,33	3, 8
	post	*Control*	16,26,27,30 34,35	1, 2, 4, 5, 7, 10, 14, 18, 22, 25, 28, 31, 32	13	6,19	8(neg), 9(neg), 11, 15, 17(neg), 20(neg)
		Treatment	1,22,26,29 32,34	5, 7, 11, 12, 14, 15, 16, 18, 19, 20, 23, 25, 27, 30, 31, 33, 35	17,28	6,9	4, 8, 10, 13(neg), 24

Discussion

Since a measurement instrument's reliability and item analysis is essential for determining individual test items' quality and utility in constructing a more reliable test, we wanted to see if the pre-test and post-test scores are reliable. The "internal structure analysis" method, one of the three main methods for estimating reliability coefficients, was used. Cronbach's alpha was calculated separately for pre-test\post-test results for each age group. The reliability estimates range from .92 to .97 for four-years-old, from .88 to .94 for five-years-olds, and from .80 to .94 for six-years-olds. While the reliability of the tests for four-year-olds is excellent, the reliability of the tests for five and six-year-olds are good.

The item analysis results identified the vocabulary test's easiest and the most difficult questions. According to the analysis, "cat", "dog" and, "spider" are the easiest items, while "seagull" is the most difficult item on the vocabulary test. Besides, according to item-test correlation results, most items` scores in the tests are strongly correlated to total test scores. A few test items are negatively correlated for a particular age and test group and are identified as the weakest items. The test can be a better measurement tool if these items are eliminated from the test.

Finally, Mixed (repeated) ANOVA statistics were performed, and three null hypotheses were tested for all age groups separately regarding interaction effect and main effects. According to Mixed ANOVA analysis, there is an interaction effect between pre-test/post-test (within-subject factor-time) results and control/treatment (between-subject factor-group) groups for four-year-olds ($p=0.00$, $p<0.005$) and five-year-olds ($p=0.045$, $p<0.05$); $F(1,19) = 19.70$, pHF=.00 and $F(1,22) = 4.51$, pHF=.00 respectively. Besides, pre-test and post-test results for all age groups are significantly different ($p=0.00$, $p<0.05$ for each age group) and correlated. Our data for all age groups are consistent with the sphericity assumption (epsilon HF=1.00).

Conclusion

Most of the studies in the literature concluded with positive outcomes about utilizing AR in education and foreign language learning (Gundogmus et al, 2016; Kucuk et al., 2014; Salmon & Nyhan, 2013; Solak & Cakir, 2017). Similarly, this experimental study's results are consistent with the current literature showing that children learn a foreign language significantly better using mobile AR apps than traditional methods. We believe that is because of the surprising factor that AR adds to the learning process and grabs children's attention.

This study focuses on the statistical analysis of the effectiveness of AR technology as a language learning tool for pre-elementary school children. The statistical analysis methodology applied in this study can provide guidance and examples for researchers working on foreign language education and willing to apply statistical methods to analyze their results. Similar studies might be conducted with different age groups (i.e., elementary school students), with different AR applications teaching different academic skills in math, science, and social sciences.

References

Antonaci, A., Klemke, R., & Specht, M. (2015). Towards design patterns for augmented reality serious games. In Springer (pp. 273-282). https://doi.org/10.1007/978-3-319-25684-9_20

Barreira, J., Bessa, M., Pereira, L. C., Adão, T., Peres, E., & Magalhães, L. (2012). MOW: Augmented reality game to learn words in different languages: Case study: Learning English names of animals in elementary school. Proceedings (pp. 1-6).

Carmigniani, J., & Furht, B. (2011). Augmented reality: An overview. In B. Furht (Ed.), Handbook of Augmented Reality (pp. 3-46). Springer.

Chang, Y. J., Chen, C. H., Huang, W. T., & Huang, W. S. (2011). Investigating students' perceived satisfaction, behavioral intention, and effectiveness of English learning using augmented reality. In Proceedings - IEEE International Conference on Multimedia and Expo (pp. 1-6).

Chen, R., & Chan, K. K. (2019). Using augmented reality flashcards to learn vocabulary in early childhood education. Journal of Educational Computing Research, 57(7), 1812-1831. https://doi.org/10.1177/0735633119854028

Dalim, C. S. C., Piumsomboon, T., Dey, A., Billinghurst, M., & Sunar, S. (2016). TeachAR: An interactive augmented reality tool for teaching basic English to non-native children. In Adjunct Proceedings of the 2016 IEEE International Symposium on Mixed and Augmented Reality, ISMAR-Adjunct 2016 (pp. 344–345).

Emmerich, F., Klemke, R., & Hummes, T. (2017). Design patterns for augmented reality learning games. In J. Dias, P. Santos, & R. Veltkamp (Eds.), Games and Learning Alliance. GALA 2017. Lecture Notes in Computer Science (Vol. 10653, pp. 161-172). Springer.

Fossey, A. (2013, December 13). Item analysis report – Item-total correlation discrimination. Retrieved from https://blog.questionmark.com/item-analysis-report-item-total-correlation-discrimination

Frey, F. (2017). Test theory, classical test theory. In International Encyclopedia of Communication Research Methods (pp. 1–6), https://doi.org/10.1002/9781118901731.iecrm0247.

George, D., & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.). Boston, MA: Allyn & Bacon.

Gundogmus, N., & Orhan, G. (2016). Foreign language teaching with augmented reality application. The Eurasia Proceedings of Educational & Social Sciences, 4, 309–312.

Ibrahim, A., Huynh, B., Downey, J., Hollerer, T., Chun, D., & O'donovan, J. (2018). ARbis Pictus: A study of vocabulary learning with augmented reality. IEEE Transactions on Visualization and Computer Graphics, 24(11), 2867–2874.

James, A., & Christensen, P. M. (2008). Research with children: Perspectives and practices (2nd ed.). New York, NY: Routledge.

Kayaoğlu, M. N., Akbaş, R., & Öztürk, Z. (2011). A small-scale experimental study: Using animations to learn vocabulary. Turkish Online Journal of Educational Technology, 10, 24–30.

Kline, T. J. (2005). Classical test theory: Assumptions, equations, limitations, and item analyses. In Psychological Testing: A Practical Approach to Design and Evaluation (pp. 91–106). SAGE Publications, Inc.

Küçük, S., Yilmaz, R. M., & Göktaş, Y. (2014). Augmented reality for learning English: Achievement, attitude and cognitive load levels of students. Egitim ve Bilim, 39, 393–404.

Malet, F., McSherry, D., Larkin, E., & Robinson, C. (2010). Research with children: Methodological issues and innovative techniques. Journal of Early Childhood Research, 8(2), 175–192.

Pellas, N., Fotaris, P., Kazanidis, I., & Wells, D. (2018). Augmenting the learning experience in primary and secondary school education: A systematic review of recent trends in augmented reality game-based learning. Virtual Reality, 329-346.

Saidin, N. F., Halim, N. D. A., & Yahaya, N. (2015). A review of research on augmented reality in education: Advantages and applications. International Education Studies, 1–8.

Salkind, N. J. (2010). Encyclopedia of research design. SAGE Publications, Inc.

Salmon, J., & Nyhan, J. (2013). Augmented Reality Potential and Hype: Towards an evaluative framework in foreign language teaching. The Journal of Language Teaching and Learning, 54-68.

Scrivner, O., Madewell, J., Buckley, C., & Perez, N. (2016). Augmented reality digital technologies (ARDT) for foreign language teaching and learning. In FTC 2016 - Proceedings of Future Technologies Conference (pp. 395-398).

Solak, E., & Cakır, R. (2017). Investigating the role of augmented reality technology in the language classroom. Croatian Journal of Education, 18.

Suruchi, S., & Rana, S. S. (2012). Test item analysis and relationship between difficulty level and discrimination index of test items in an achievement test in Biology. Paripex Indian Journal of Research, 3, 56-58.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's Alpha. International Journal of Medical Education, 2, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Vate-U-Lan, P. (2012). An Augmented Reality 3D Pop-Up Book: The development of a multimedia project for English language teaching. Proceedings of the 2012 International Conference on Multimedia Engineering (pp. 890-895). https://doi.org/10.1109/ICME.2012.79

Yilmaz, R. M. (2016). Educational magic toys developed with augmented reality technology for early childhood education. Computers in Human Behavior, 54, 240–248.

Zeng, J., & Wyse, A. (2009). Introduction to classical test theory. Paper presented at the Michigan Department of Education; Office of Educational Assessment and Accountability. Retrieved from https://www.michigan.gov on January 21, 2023.